DOCUMENT RESUME

ED 453 252 TM 032 789

AUTHOR Lee, Jaekyung; Coladarci, Theodore
TITLE Imperative or Choice? Multi-Level and Multi-Measure Analysis of Student Assessment Data for Evaluation of Systemic School Reform.
SPONS AGENCY National Science Foundation, Washington, DC.
PUB DATE 2001-04-11
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; *Educational Change; Elementary Secondary Education; Estimation (Mathematics); *Evaluation Methods; *School Effectiveness; State Programs; Testing Programs
IDENTIFIERS *Kentucky; *Maine; Multilevel Analysis; National Assessment of Educational Progress; Reform Efforts; Systemic Educational Reform; Weighting (Statistical)

ABSTRACT

A systematic analysis of student assessment data from Maine and Kentucky, using National Assessment of Educational Progress (NAEP) data and state and local assessment results, was conducted to address issues of measurement and attribution involved in evaluating systemic school reform. This paper (1) examines ways to cope with the challenges of considering measures of school systems from multiple sources and combining multiple measures of student achievement data using state and local data from Maine; (2) examines ways to tackle the challenges of considering multiple levels of influences on student achievement and attributing achievement results to school effects using NAEP data from Maine and Kentucky; and (3) discusses the usefulness and limitations of multi-level and multi-measure approaches to the evaluation of systemic school reform. Results suggest that it is not necessary to weight each measure before forming an achievement composite to classify student performance. This is particularly true when measures are highly intercorrelated. Results also point to the possible hazards of classifying student achievement based on a single measure. Three different models of estimating school effects were tested, and reasons for their use are discussed. The estimation of school effects requires that "school effects" be defined so that an explicit model of these effects can be formulated. The model should be fully specified, with all variables representing school input, practice, context, and student background measured. (Contains 14 tables and 17 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 252

Imperative or choice?

Multi-level and multi-measure analysis

of student assessment data for evaluation of systemic school reform

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Jaekyung Lee Theodore Coladarci

College of Education and Human Development

University of Maine

Paper presented at the 2001 meeting

of the American Educational Research Association, Seattle

TM032789

We acknowledge that this research has been supported by a grant from the National Science Foundation. The views expressed herein are solely those of the authors.

BEST COPY AVAILABLE

1. Research Objectives

During the last decade, many states have initiated systemic school reform. Systemic school reform is aimed at improving academic excellence for all students at all levels of the school system simultaneously (Smith & O'Day, 1991). Evaluation of systemic school reform calls for coordinated collection of information on student achievement at the different levels of school system (Roeber, 1995). At the same time, accountability piece of systemic school reform requires value-added school performance indicators. These policy imperatives lead us to investigate the adequacy and utility of methods for assessing and understanding the performance of a school system involved in systemic school reform.

In light of these concerns, we conduct a systematic analysis of student assessment data from Maine and Kentucky—the National Assessment of Educational Progress (NAEP) and state and local assessments—to address the issues of measurement and attribution involved in evaluating systemic school reform. This paper consists of three major sections. First, we examine ways to cope with the challenges of considering measures from multiple sources of school system and combining multiple measures of student achievement data (measurement issue). For this analysis, we use state and local assessment data collected in Maine. Second, we examine ways to tackle the challenges of considering multiple levels of influences on student achievement and attributing achievement results to school effects (attribution issue). For this analysis, we use NAEP data collected in Maine and Kentucky. Third, we discuss the utility and limitations of multi-level and multi-measure approaches to evaluation of systemic school reform.

2. Combining Multiple Measures of Achievement

A number of state and federal agencies now recommend or require multiple measures to assess student achievement (Ardovino, Hollingsworth, & Ybarra, 2000). However, no criteria about reliability, validity, and weighting in using multiple measures have been set by states like California (Jang, 1998). Currently available measures of student achievement are often inadequate for evaluation of systemic school reform, particularly when they rely on norm-referenced standardized tests and use percentile ranks as grade level standards. While local assessments are a potentially valuable source of additional measures, there is often insufficient consistency of the measures across sites. Despite these problems and challenges, districts have devised their own ways to combine multiple measures of achievement, which produces a great deal of variation from district to district (see Jang, 1998; Kalls, 1998; Law, 1998; Novak, Winters, & Flores, 2000).

In the present climate of standards-based education, school leaders in Maine also are being asked to think about assessment in new ways. Student achievement of the state standards, the *Learning Results*, must be measured by a combination of state and local assessments. Based on these assessments, local educators soon will be expected to "certify" a student's attainment of the *Learning Results* in order for the student to receive a high school diploma.

How should we approach the challenge of combining multiple measures of achievement for arriving at a single judgment of, say, "proficiency," or "meeting the

standard"? Specifically, what is an efficient and defensible method for combining multiple measures of achievement? This is the general question that we address in this section.

Data collection and analysis

We collected data from two sites in Maine, which were chosen because of their similarity in community size and proximity to the University of Maine. In both sites, we obtained the following achievement information for each student: (a) the mathematics subscale score on the 8th grade Maine Educational Assessment (MEA-M), (b) the mathematics subscale score on the locally administered standardized achievement test (ITBS in Site A and TerraNova in Site B), and (c) the course grade achieved in mathematics. In Site A ($n = 94$), all information was taken in the student's 8th grade year; in Site B ($n = 65$), the standardized achievement test and mathematics grades were obtained in the 9th grade (see Table 1). The MEA-M scores provide the only truly meaningful achievement information for comparing the two sites. From Table 2, one sees that the MEA-M mean for Site B was 17.76 points higher than that for Site A. With a pooled within-group standard deviation of 15.77, this mean difference corresponds to an effect size of $d = 17.76 \div 15.77 = +1.13$.

Creating a Common Scale for Mathematics Course Grade

As can be seen from Table 1, students in each site did not all enroll in the same level of mathematics. Our first task, then, was to create a single variable for "mathematics grade," even though it would comprise grades from different classes. Although we followed the same procedure in both sites, we will illustrate this procedure using data from Site A.

Site A students received a grade, on a 100-point scale, for either general mathematics (n = 59), algebra 1 (n = 29), or geometry (n = 6) (see Table 3). Because we believe that it makes little sense to regard a final grade in general mathematics as being comparable to the same grade in a higher level class, we weighted algebra 1 and geometry grades according to how these two groups of students performed on the MEA-M relative to the general mathematics students (see Table 4). Each of the two mean differences was converted to an effect size:

$$d_{21} = \frac{531.72 - 514.64}{9.46} = +1.81$$

$$d_{31} = \frac{555.00 - 514.64}{9.46} = +4.27$$

where d_{21} represents the difference in MEA-M scores between student enrolled in algebra 1and those taking general mathematics, and d_{31} the difference in MEA-M scores between geometry students and those taking general mathematics. Each effect size was then used to adjust upwards the mathematics grades for students enrolled in either algebra 1 or geometry. We did this by multiplying the pooled within-group standard deviation for mathematics grades (8.31) by either d_{21} or d_{31}, and then adding the product to the student's math grade. This resulted in an adjustment of +15.04 for each of the 29 algebra 1 students and +35.49 for the 6 geometry students. The resulting scale, which pools the three mathematics classes, is $\overline{X}$ = 89.24 and SD = 17.65.

Analyses and Results

Correlational Analyses

To examine the relationships among the results of state and local assessments, we obtained student-level within-site correlations among the three measures of student

achievement: (a) MEA-M, (b) the mathematics subscale score on the locally administered standardized achievement test (which we refer to as "ITBS/TN"), and (c) the weighted course grade achieved in mathematics ("COURSE").

As Table 5 shows, the three measures of mathematics achievement correlate substantially. Although these correlations are uniformly high, there is some variation in magnitude. Interestingly, COURSE correlates more highly with MEA-M than with ITBS/TN. This is not surprising, insofar as one would expect classroom assessments and the MEA to align with the *Learning Results* more than would be expected of a commercially available standardized test.

Classification Analyses

To explore an efficient and defensible method for combining multiple measures of achievement, we combined the three measures two different ways and compared the results by conducting classification analyses. As with the correlational analyses, these analyses were conducted within site.

Because of the standard setting process that was employed in the development of the Maine Educational Assessment, MEA-M scores can be stated in terms of performance levels that are tied to state standards:

exceeds the standard:	561
meets the standard:	541
partially meets the standard:	521
does not meet the standard:	<521

The critical score here is 541 (on a scale of 501-580), which is the cutscore that distinguishes between meeting the standard and not.

Although Maine school leaders soon will be expected to engage in standard

setting for their local assessments, the two sites in the present study, like most Maine school districts, have yet to implement standard setting. Consequently, neither COURSE nor ITBS/TN can be directly expressed as a performance level within the context of the *Learning Results*. However, because MEA-M correlates highly with both ITBS/TN and COURSE (Table 5), we can estimate, using simple regression, the critical cutscore for each of the latter two measures. We began by regressing ITBS/TN on MEA-M and, given the resulting equation, determined the predicted value of ITBS/TN for MEA-M = 541 (i.e., the designated cutscore for "meets the standard"). In Site A, for example, this regression equation is:

$$\text{ITBS/TN} = -676.487 + 1.4(\text{MEA-M})$$

which, for MEA-M = 541, yields an estimated cutscore of 80.91 (in percentile rank) for ITBS/TN. The analogous procedure was followed for COURSE. Again, for Site A this equation is:

$$\text{COURSE} = -443.307 + 1.019(\text{MEA-M})$$

which yields an estimated cutscore of 107.97 (in weighted grade) for COURSE. Thus, we identified in each site the score for ITBS/TN and for COURSE that corresponds to the MEA-M threshold for meeting the state standard.

We then transformed MEA-M, ITBS/TN, and COURSE to *z*-scores using the standard formula, but with one modification: We replaced the mean with 541 in the transformed MEA-M variable and the estimated cutscore (as described above) in the transformed COURSE and ITBS/TN variables. With this substitution, the sign of a *z*-score now indicates the student's performance relative to the MEA-M standard (rather than to the parent variable's mean).

Next, we formed an *unweighted* composite by taking the simple mean of the three transformed variables. A negative value on this composite went to the student who, on average, fell below the "standard" on all three measures. We also formed a *weighted* composite by (a) subjecting the three measures to a principal components analysis and (b) using the resulting component score coefficients to weight each measure in the formation of the composite. Each composite was dichotomized at 0, as were the transformed MEA-M, COURSE, and ITBS/TN variables. We then examined classification similarity by constructing a series of 2 x 2 tables.

The fundamental question is whether the unweighted and weighted composites classified students similarly. That is, when forming an achievement composite, is anything gained by weighting the measures that enter into the composite? As Table 6 shows, there was perfect agreement between the two sets of classifications. This no doubt reflects the relatively uniform correlations among MEA-M, ITBS/TN, and COURSE (Table 5) and, in turn, the relatively uniform component score coefficients that we obtained from the principal components analysis (see Table 7). In short, the results of this analysis indicate that weighting each measure is unnecessary. Thus, if the choice is between weighting or not weighting, the most efficient strategy for combining multiple measures would appear to be the latter. This assumes that correlations among measures are similar (which should be examined empirically) and that the measures are of equal importance. If either assumption does not hold, then weighting would be defensible.

A secondary question concerns the level of agreement between the classification based on the unweighted composite and that based on a single measure (see Tables 8-10). Except for the perfect agreement in Site A involving MEA-M, the levels of agreement are

fairly consistent, ranging from 89% to 92%. In these later cases, single-measure classification resulted in more students meeting the standard than when classification was based on the composite.

3. Identifying School Effects on Achievement

Student achievement is critically affected by variables at different levels of school organization. If academic achievement depends on the characteristics of students and teachers and/or the organizational context in which teaching and learning occurs, one cannot meaningfully assess school effects without considering these multi-level sources of influences (Keeves & Sellin, 1988). Previous studies of school effects in Maine and Kentucky analyzed aggregate school data to examine variation among schools in their performance status and gain, and found that poverty was the strongest and most consistent predictor of school performance in both states (Lee, 1998; Roeder, 2000). The past school performance indicators tend to focus on average test scores, which possibly conceal achievement differences among groups of students within each school. Consequently, these analyses are not sensitive to equity-related issues. Even when the effects of student-level background characteristics on achievement were considered to estimate value-added school performance, the effects are often assumed to be uniform across schools.

Multilevel analysis methods not only provide a means for formulating student-level and school-level regression models simultaneously, but they also provide more precise estimates of the relationships between predictors and outcomes at each level

(Bryk & Raudenbush, 1992). In particular, hierarchical linear modeling (HLM) is popular among educational researchers and evaluators for estimating school effects (see Phillips & Adcock, 1997; Weerasinghe, Orsak, & Mendro, 1997; Yen, Schafer, & Rahman, 1999). Because public schools do not randomly assign students and teachers across schools, multilevel methods that account for student and school context variables are regarded as the most rigorous means for estimating school effects (Phillips & Eugene, 1997). In fact, HLM has been found to produce more stable school effect estimates than ordinary least squares (OLS) or weighted least squares (WLS) methods (Yen et. al., 1999). This is true particularly when schools have few students and, thus, OLS estimates of the within-school regression parameter have low reliability.

Raudenbush and Willms (1995) discuss two different types of school effects: Type A and Type B effects. Type A effect is the difference between a child's actual performance and the expected performance had that child attended a typical school. This effect doesn't concern whether that effectiveness derives from school inputs (e.g., class size, teacher quality) or from factors related to school context (e.g., community affluence, parental support). By contrast, a Type B effect isolates the effect of a school's input from any attending effects of school context. The two indicators are appropriate for purposes of school choice and school accountability, respectively (Meyer, 1997). When HLM methods have been used to obtain school effect indices, researchers often did control for the influences of student background variables. However, the corresponding school-level compositional effects of these variables were not taken fully into account (see Weerasinghe, Orsak, & Mendro, 1997; Yen, Schafer, & Rahman, 1999). Raudenbush and Willms (1995) also suggest considering the possibility that a school will influence

different students differently. Yet there has been little research that systematically examines the achievement gaps among different groups of students as school effect indices.

How should we approach the challenge of identifying value-added contribution of schools to academic achievement for arriving at a judgment of, say, "effective"? Specifically, what is an efficient and defensible method for determining school effectiveness? This is the general question that we address in this section.

Data and Methods

In the present study, we use the data collected under 1996 NAEP 8th grade state math assessments for Kentucky and Maine. This allows us to compare the two states in terms of their school effects. The NAEP data are hierarchical in nature because students are nested within schools. HLM addresses the problem of students nested within schools. Further, the use of HLM on NAEP data copes with the problem of sampling error resulting from the multi-stage sampling in NAEP (see Arnold, 1993). Using HLM, we examine the effects of race and socioeconomic status on achievement at the student and school levels to estimate (a) adjusted school average achievement and (b) within-school racial and social gaps in achievement. We also examine relationships among the school performance indices obtained from HLM separately in each state. Finally, we compare schools in Maine and Kentucky from pooled HLM analyses and discuss implications of their differences for school effectiveness research.

Taking a multi-level organizational perspective and drawing on the relevant literature, we test three models of school effects separately for Maine and Kentucky: Model 1 (no predictors at the student and school levels), Model 2 (predictors at the

student level only, with grand-mean centering), and Model 3 (predictors both at the student and school levels, with grand-mean centering). Type A effect is estimated through Model 2 by removing the effect of student background variables. Type B effect is estimated through Model 3 by removing the effects of variables beyond a school's control (e.g., demographic composition). In this study, we consider only race and SES (socioeconomic status) factors. We believe that students' prior achievement (readiness for learning measured at the time of entry into current school) and mobility (length of stay in current school) factors must be considered to estimate authentic school effects but these data are not available in the NAEP.

All analyses were conducted using the HLM 5 program. Table 11 presents descriptive statistics for all variables used in these analyses. MRPCM1 through MRPCM5 are the five plausible values that make up the composite mathematics achievement outcome variable. WHITE is a dummy variable (1 = white, 0 = minority), and SES is a composite factor of parental education level, availability of reading materials at home, and school median income (standardized to have a mean of 0 and a standard deviation of 1 across states).

Model 1

Model 1, which includes no predictors at the student and school levels, partitions the total variance in mathematics achievement into its within- and between-school components. The school-level residual value from this model is used as an indicator of unadjusted school average performance.

Model 2

Model 2 adds student-level predictors by regressing mathematics achievement for student i within school j on race (WHITE) and socioeconomic status (SES). The Level 1 model (student level) is

$$(MRPCM)ij = \beta_{0j} + \beta_{1j}(WHITE)_{ij} + \beta_{2j}(SES)_{ij} + e_{ij}$$

where (MRPCM)ij is the composite mathematics achievement of student i in school j; $(WHITE)_{1ij}$ is the indicator of student i's race in school j; (SES)ij is the indicator of student i's socioeconomic status in school j; and e_{ij} is a Level 1 random effect representing the deviation of student ij's score from the predicted score based on the student-level model. Level 1 predictors are grand-mean centered so that the intercept, β_{0j}, can be interpreted as adjusted mean achievement for school j. This adjustment is chosen to sort out the unique effects of school on achievement after controlling for the influences of student/family characteristics.

The next step in HLM involves fitting an unconditional, or random, regression model at the school level (Level 2). Notice that all Level 1 regression coefficients are regarded as randomly varying across schools, and γ_{00} is the mean value of the school-level achievement outcome beyond the influences of student/family characteristics. r_{0j}, the school-level residual value from this regression, is used as an indicator of school average performance adjusted for racial and SES mixes of students. Likewise, r1j and r2j are used as indicators of racial and social achievement gaps respectively. The Level 2 (school level) model is

$$\beta_{0j} = \gamma_{00} + r_{0j}$$

$$\beta_{1j} = \gamma_{10} + r_{1j}$$

$$\beta_{2j} = \gamma_{20} + r_{2j}$$

where β_{0j} represents school j's average mathematics achievement adjusted for its composition of students' racial and SES backgrounds; β_{1j} represents school j's racial gap (i.e., the achievement score gap between white and minority students); and β_{2j} represents school j's social gap (i.e., the extent to which students' SES differentiates their achievement).

Model 3

Model 3 adds two school-level predictors, or, school aggregate values of student-level predictors. Percent white (PWHITE) and average SES (AVSES) are added to explain between-school variation. r_{0j}, the school-level residual value from this regression, is used as an indicator of school average performance adjusted for racial and social composition effects. Model 3 is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(PWHITE)_j + \gamma_{02}(AVSES)_j + r_{0j}$$

where $(PWHITE)_j$ is the proportion of white students (i.e., the mean of WHITE) in school j; and $(AVSES)_j$ is the mean SES of school j.

Results

Model 1 (fully unconditional model)

Decomposition of variance in the outcome variable shows that the two states have similar distributions of mathematics achievement between the school and student levels.

In Maine,18% of variance exists at the school level and 82% at the student level; the figures are 17% and 83%, respectively, in Kentucky. Residual school means from this model are called Model 1 average. The reliability estimate of these unadjusted school achievement averages is .80 in Maine and .79 in Kentucky, indicating that the sample means tend to be quite reliable as indicators of the true school means.

Model 2 (level-1 predictors only with grand-mean centering)

By using race and SES variables as predictors of math achievement at the student level (with grand-mean centering), we obtain adjusted school average achievement that takes into account differences among schools in their students' racial and social mixes. A residual school mean that is obtained after controlling for the effects of student-level predictors, as an indicator of value-added school performance, is called Model 2 average. The reliability of conditional school means (conditional reliability) becomes lower: .67 in Maine and .62 in Kentucky. As shown in Table 12, Model 2 average is correlated very highly with Model 1 average (r_{me}=.92 and r_{ky}=.87).

The effects of race and SES on achievement are used as indicators of academic inequity, as well as providing the basis for adjusting estimates of school effects. This assumes heterogeneity of regressions among schools and models the effects of student's race and SES on achievement as randomly varying at the school level. The within-school racial gap—the estimated average achievement gap between white and minority students within schools—is 12.1 (.41 standard deviations) in Maine and 16.8 (.57 *SD*) in Kentucky (see Table 13). The within-school social gap—the estimated effect of SES on achievement within schools—is 10.8 (.38 *SD*) in Maine and 10.6 (.36 *SD*) in Kentucky (see Table 13). In both states, these gaps are highly significant.

Maine and Kentucky show different patterns of relationships between achievement average and gap estimates (Table 12). In Maine, Model 2 average correlates positively with racial gap (.72) but negatively with social gap (-.63). Conversely, in Kentucky, Model 2 average correlates negatively with racial gap (-.28) but positively with social gap (.57). Higher performing schools in both states tend to have smaller gaps with regard to one background variable but larger gaps with regard to the other. This indicates that schools are not very effective in addressing both racial and social achievement gaps.

We should note that the reliability estimates of racial and social gaps are low: .13 and .21 in Maine, and .30 and .28 in Kentucky. Considering these reliabilities, it appears that both Maine and Kentucky schools vary little in their racial and social gaps. This is attributed to the fact that both states are highly homogeneous in racial composition. However, sufficient variability across schools on racial gap estimates does exist as the homogeneity of variance tests demonstrate significant variation (see the variance component chart in Table 13).

Model 3 (both level-1 and level-2 predictors with grand-mean centering)

School-level predictors of racial and social composition were used to make further adjustment for differences among schools in their average achievement due to composition effects. In Maine, both racial and social composition effects are not significant. This indicates that such school-level adjustment of performance for race and SES factors, in addition to the corresponding student-level adjustment, is not necessary (see Table 13). In Kentucky, only the social composition effect is significant, adding about 7 points to the within-school social gap estimate (see Table 13). Model 3

average—residual school means after controlling for both student and school-level effects of race and SES—correlates .70 with Model 1 average and .94 with Model 2 average (see Table 12).

Pooled HLM analysis

In order to test differences in school performance between Maine and Kentucky, we pooled data from the two states and applied the same three models. However, we added a school-level dummy variable (MAINE) to indicate where a school's location (Maine = 1, Kentucky =0).

The results of the pooled HLM analyses are summarized in Table 4. First, the comparison of Maine and Kentucky schools without any control for background variables show that Maine schools perform significantly better than Kentucky schools: a gap of 17.18 (Model 1), or roughly 1.2 *SD*. The gap between Maine schools and their Kentucky counterparts in terms of average 8th grade mathematics achievement decreases about 40% when we control for their differences in students' racial and social background variables (gap = 9.97, Model 2). When we further control for school composition effects, the Maine-Kentucky school achievement gap becomes slightly smaller but remains statistically significant (gap = 6.18, Model 3). As Maine schools turn out to perform significantly better than Kentucky schools based on both Type A and Type B effect estimates, their effectiveness gap seems to come from sources related to schooling; students' prior achievement and mobility factors become less important when we compare schools across states (vs. within state). Despite the average school performance gap, it turned out that there are no significant differences between the two states' schools in terms of their racial and social gap estimates.

4. Discussion

Evaluation of systemic school reform requires that we evaluate school performance with multiple measures at mutliple levels of school system. This policy imperative makes data collection and analysis very challenging and complex. Despite the imperative, there is a lot of room for us to make technical choices that must be informed by scientific research. Although our results may not generalize to all states, they are expected to inform us about desired data and methods for a more systematic evaluation of systemic school reform. We caution that analytical methods themselves cannot cope with inherent measurement and attribution problems. We discuss implications of our research findings below.

Multi-measure Analysis of Student Achievement

Our results suggest that it is not necessary to weight each measure before forming an achievement composite to classify student performance. This is particularly true where measures are highly intercorrelated, as was the case here. If intercorrelations vary in magnitude, however, then it may be advisable to weight each measure to reflect the measure's association with the underlying principal component. Subsequent research would throw clarifying light on the merits of this recommendation, especially if the research involves multiple sites that differ with respect to the relatedness of the achievement measures they employ.

Having said this, we should acknowledge that high intercorrelations among measures are not sufficient for deciding in favor of an unweighted composite. That is,

one also should take into account the announced importance of each measure. For example, if a school district attaches greater importance to a district-wide assessment compared to, say, the standardized test that is annually administered, then the former should receive greater weight—even in the face of a high correlation between the two. Although there are various reasons why local achievement measures may differ in importance, a primary reason is the degree to which a measure aligns—in various respects (e.g., see Webb, 1997)—with the adopted standards. The reliability of assessment measures also need to be considered in developing weights.

Our results also point to the possible hazards of classifying student achievement based on a single measure. As Tables 8-10 illustrate, single-measure classification tended to result in additional students identified as meeting the standard. Are these students false positives? Because of two limitations of the present study, we unfortunately do not know. First, unlike MEA-M, which was designed to align with the *Learning Results*, neither ITBS/TN nor COURSE was constructed explicitly to reflect student attainment of these standards. This clearly is true for ITBS/TN, for no commercially available standardized achievement test is tailored to the standards of a particular state. And although teacher-constructed mathematics assessments (COURSE) in Maine arguably are more responsive to the *Learning Results*, the task of formally designing classroom assessments to demonstrably align with these standards still looms on the horizon for most Maine school districts. Clearly, in a standards-based climate, the integrity of an achievement composite depends, in part, on the extent to which the component measures are drawing on the same universe of standards. Without this assurance, we must interpret with caution the tendency of the single-measure

classifications to putatively overidentify students who meet the standard. Here, too, subsequent research could be illuminating, particularly if the research involves multiple sites that vary with respect to the degree to which each measure is of demonstrable alignment with the announced standards.

A second, and related, limitation of the present study is that neither site had engaged in formal standard setting for either ITBS/TN or COURSE—hence our decision to obtain regression estimates of ITBS/TN and COURSE cutscores, given the relationship between each measure and the MEA-M (for which the minimum score for "meets the standard" is known).

Multilevel Analysis of School Effects

We have tested three different models of estimating school effects. Model 2 is regarded as fairer than Model 1 as it considers student background factors that schools cannot control. Model 3 also may be fairer than Model 2 as it further takes into account school-level compositional effects beyond individual student-level effects and implies comparing "like with like." However, this position can be challenged in a situation where there is systematic covariation between school context and school practice variables. Raudenbush and Willms (1995, p. 332) point out the problem of causal inference:

> "Causal inference is much more problematic in the case of Type B effects because the treatment—school practice—is typically undefined so that the correlation between school context and school practice cannot be computed. Thus, even if the assignment of students to schools were strongly ignorable, the assignment of schools to treatments could not be."

Bryk and Raudenbush (1992, p.128) illustrate the problem where there exists differences in school staff quality that might confound the effects of school staff with the effects of student composition:

> "Suppose that [high SES] schools have more effective staff and that staff quality, not student composition, causes the elevated test scores. The results could occur, for example, if the school district assigned its best principals and teachers to the more affluent schools. If so, [Model 3] would give no credit to these leaders for their effective practices."

Conversely, one might argue that the differences among schools in school resources (including class size, teacher/administrator quality and instructional resources), possibly due to their different student demographic composition, are precisely what we need to remove for evaluating schools in fair ways. If high SES schools do a better job simply because they draw better staff, more resources, and better students, then this advantage should not be considered authentic "school" effects—i.e., differences among schools due to educational efforts and practices. Then, the task becomes to distinguish school inputs that are determined outside the school and sort out their effects as external school-level characteristics (Meyer, 1997). But this strategy can be more problematic when the school input variables are more highly correlated with school practice variables.

Thus, the fundamental issue is not simply a technical choice of estimation methods given the available data. Rather, the estimation of school effects requires that we define "school effects" and formulate an explicit model of these effects. In other words, this approach requires that the model be fully specified: all variables representing school input, practice, context, and student background would have to be measured and

included in the model in order to guarantee that the effects of school practice were unbiased. Nevertheless, school quality variables are generally more difficult to define and measure and the relevant data are expensive to collect (Raudenbush & Willms, 1995).

Our analysis of school effects also involved estimating student achievement gaps with regard to background characteristics (i.e., race and SES in our case). We found that while average achievement varies significantly among schools in both states, their racial and social gaps vary little among schools. This means that much of the observed variability in achievement gaps is sampling variance and, as a result, cannot be explained by school factors. Thus, at least in our data, it is not sensible to use student achievement gaps as school effect indices. It remains to be seen whether combination of state and local assessment measures would produce different results than those based on the NAEP.

References

Ardovino, J., Hollingsworth, J., & Ybarra, S. (2000). Multiple measures: Accurate ways to assess student achievement. Thousand Oaks, CA: Corwin Press.

Arnold, C. A. (1993). Using HLM with NAEP. Unpublished Paper Presented at the Advanced Studies Seminar on the Use of NAEP Data for Research and Policy Discussion, Washington, D.C.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models. Newbury Park: Sage Publication.

Jang, Y. (1998). Implementing standards-based multiple measures for IASA, Title I accountability using Terra Nova multiple assessment. Paper presented at the annual meeting of the AERA. (ED 426 084).

Keeves, J. P., & Sellin, N. (1988). Multilevel analysis. In J. P. Keeves (Ed.) Educational research, methodology, and measurement: An international handbook. New York: Pergamon Press.

Kolls, M. R. (1998). Standards-based multiple measures for IASA, Title I program improvement accountability: A vital link with district core values. Rowland unified school district. Paper presented at the annual meeting of the AERA. (ED 420 681).

Law, N. (1998). Implementing standards-based multiple measures for IASA, Title I accountability using Sacramento achievement levels. Paper presented at the annual meeting of the AERA. (ED 421 497).

Lee, J. (1998). Assessing the performance of public education in Maine: A national comparison. Orono, ME: University of Maine Center for Research and Evaluation.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. Economics of Education Review, 16(3), 283-301.

Novak, J. R., Winters, L., Flores, E. (2000). Using multiple measures for accountability purposes: one district's experience. Paper presented at the annual meeting of the AERA. (ED 443 846).

Phillips, G. W., & Adcock, E. P. (1997). Measuring school effects with HLM: data handling and modeling issues. Paper presented at the annual meeting of the AERA. (ED 409 330).

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statistics. 20(4), 307-335.

Roeber, E. (1995). Emerging student assessment system for school reform. ERIC Digest (ED 389 959).

Roeder, P. W. (2000). Education reform and equitable excellence: The Kentucky experiment. Unpublished research paper.

Weerasinghe, D., Orsak, T., Mendro, R. (1997). Value added productivity indicators: A statistical comparison of the pre-test/post-test model and gain model. Paper presented at the annual meeting of the Southwest Educational Research Association. (ED 411 245)

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 6. National

Institute for Science Education (NISE), University of Wisconsin—Madison. Washington, DC: NISE.

Yen, S., Schafer, W. D., & Rahman, T. (1999). School effect indices: stability of one- and two-level formulations. Paper presented at the annual meeting of the AERA. (ED 430 029).

Table 1.
When achievement information was collected, by site.

achievement information ↓	*Site A* (n = 94)	*Site B* (n = 65)
Maine Educational Assessment (mathematics score)	8th grade	8th grade
Standardized achievement test, mathematics	8th grade (Iowa Test of Basic Skills; percentile ranks)	9th grade (Terra Nova; scaled scores)
course grade, mathematics	8th grade (course grade in general math, algebra 1, or geometry)	9th grade (course grade in applied math 1,integrated math, practical math 1, algebra 1, or geometry)

Table 2.
Distribution of MEA-M mathematics scores in each site.

	MEA-M performance	
course	*M*	*SD*
Site A (n = 94)	522.49	14.88
Site B (n = 65)	540.25	16.97
	SD_{pooled} =	15.77

Table 3.
Distribution of unweighted mathematics grades for each of three courses (Site A).

course	*M*	*SD*
general mathematics ($n = 59$)	78.24	9.26
algebra 1 ($n = 29$)	88.17	6.58
geometry ($n = 6$)	94.33	4.50
		$SD_{pooled} = 8.31$

Table 4.
Distribution of MEA-M mathematics scores for students in each of three mathematics courses (Site A).

	MEA-M performance	
course	*M*	*SD*
general mathematics ($n = 59$)	514.64	9.02
algebra 1 ($n = 29$)	531.72	10.82
geometry ($n = 6$)	555.00	5.33
		$SD_{pooled} = 9.46$

Table 5.
Correlations among measures of student achievement in mathematics.

	Site A	
	MEA-M	ITBS/TN
ITBS/TN	.81	
COURSE	.86	.72
	Site B	
ITBS/TN	.85	
COURSE	.84	.77

Table 6.
Classification similarity: unweighted and weighted composites.

Site A

		weighted composite	
		below standard	meets standard
unweighted composite	below standard	82	
	meets standard		12

Site B

		weighted composite	
		below standard	meets standard
unweighted composite	below standard	33	
	meets standard		32

Table 7.
Component score coefficients.

	Site A	Site B
MEA-M	.389	.389
ITBS/TN	.368	.376
COURSE	.354	.346

Table 8.
Classification similarity: Unweighted composite and MEA-M.

Site A
(100% agreement)

		MEA-M		
		below standard	meets standard	row total
unweighted composite	below standard	82		82
	meets standard		12	12
	column total	82	12	94

Site B
(92% agreement)

		MEA-M		
		below standard	meets standard	row total
unweighted composite	below standard	29	4	33
	meets standard	1	31	32
	column total	30	35	65

Table 9.
Classification similarity: Unweighted composite and ITBS/TN.

Site A
(91% agreement)

		ITBS/TN		
		below standard	meets standard	row total
unweighted composite	below standard	75	7	82
	meets standard	1	11	12
	column total	76	18	94

Site B
(91% agreement)

		ITBS/TN		
		below standard	meets standard	row total
unweighted composite	below standard	29	4	33
	meets standard	2	30	32
	column total	31	34	65

Table 10.
Classification similarity: Unweighted composite and COURSE.

Site A
(90% agreement)

		COURSE		
		below standard	meets standard	row total
unweighted composite	below standard	75	7	82
	meets standard	2	10	12
	column total	77	17	94

Site B
(89% agreement)

		COURSE		
		below standard	meets standard	row total
unweighted composite	below standard	28	5	33
	meets standard	2	30	32
	column total	30	35	65

Table 11.

Descriptive statistics of predictors and outcome variables for HLM analyses of Kentucky and Maine 1996 NAEP 8th grade math data

	Kentucky			Maine		
	n	*M*	*SD*	*n*	*M*	*SD*
			Student-level			
MRPCM1	2461	267.29	30.88	2258	285.22	30.51
MRPCM2	2461	267.14	31.00	2258	285.89	30.19
MRPCM3	2461	266.85	30.99	2258	284.95	30.17
MRPCM4	2461	267.01	30.87	2258	284.73	30.04
MRPCM5	2461	267.25	30.78	2258	285.11	30.32
WHITE	2535	0.87	0.33	2309	0.95	0.22
SES	2230	-0.40	0.94	2103	0.17	0.83
			School-level			
PWHITE	101	0.87	0.16	93	0.95	0.06
AVSES	101	-0.42	0.52	93	0.14	0.45

Table 12.

Correlations among school performance indicators

	Model 1 average	Model 2 average	Model 3 average	Racial gap
Model 2 average	0.87			
	0.92			
Model 3 average	0.70	0.94		
	0.82	0.97		
Racial gap	-0.24	-0.28	-0.23	
	0.61	0.72	0.77	
Social gap	0.34	0.57	0.53	-0.50
	-0.52	-0.64	-0.68	-0.96

Note. Upper values are for Kentucky and lower values are for Maine.

Table 13.

Summary of HLM Results

	Kentucky		Maine	
	Model 2	Model 3	Model 2	Model 3
	Estimation of Regression Coefficients (Fixed Effects)			
School-level Effects				
Adjusted Mean Outcome	266.58***	267.29***	283.92***	283.74***
PWHITE		-.39		38.01
AVSES		7.15**		3.27
Student-level Effects				
WHITE	16.79***	16.79***	12.11***	12.11***
SES	10.58***	10.58***	10.78***	10.78***
	Estimation of Variance Components (Random Effects)			
Adjusted Mean Outcome	90.39***	81.57***	91.86***	81.90***
WHITE	141.66***	141.66***	72.60**	72.60**
SES	21.42	21.42	16.50	16.50
	Percent of Outcome Variance Explained			
school-level	38.4	44.0	37.7	44.5
student-level	15.5	15.5	9.2	9.2

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 14.

Summary of Pooled HLM Results

	Model 1	Model 2	Model 3
	Estimation of Regression Coefficients		
School-level Effects			
Adjusted Mean Outcome	266.19***	270.29***	283.92***
MAINE	17.18***	9.97***	6.18**
PWHITE			4.41
AVSES			6.72***
Student-level Effects			
WHITE		16.77***	17.01***
SES		10.52***	10.02***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$